Journal of Pioneering Medical Sciences
Received: May 10, 2025 | Accepted: June 11, 2025 | Published: August 05, 2025
Volume 14, Issue 07, Pages 77-86

DOI https://doi.org/10.47310/jpms2025140714

Immunogenicity and Structural Effects of INS Gene Mutations in Type 1 Diabetes Mellitus

Wasan Abdulateef Majeed[*1], Ola A. Kadhim[2] and Eman Salman Khamaes[3]
[1]Department of Biology, College of Education for Pure Science, Iraq
[2]Department of Biotechnology, College of Science, University of Anbar, Iraq
[3]Department of Microbiology, College of Medicine, University of Diyala, Diyala, Iraq

Author Designation: [1,2]Lecturer, [3]Associate Lecturer

*Corresponding author: Wasan Abdulateef Majeed (e-mail: wassan.abdullateef@uodiyala.edu.iq).

©2025 the Majeed, Wasan Abdulateef *et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0

Abstract Type 1 Diabetes (T1D) is an autoimmune disease driven by the destruction of insulin-producing beta cells. Genetic susceptibility includes variations in the *INS* gene, which encodes insulin. We hypothesized that T1D-associated *INS* mutations may alter insulin's structure, affect its immunogenicity and contribute to autoimmunity. To investigate this, we conducted the first comprehensive structural analysis of all T1D-linked *INS* missense variants reported in ClinVar, using a multi-faceted computational approach. We curated all reported INS missense variants linked to T1D from ClinVar and analyzed their impact using Molecular Dynamics (MD) simulations. Leveraging the InterPro database, we mapped PR proinsulin's domain architecture and assessed evolutionary conservation via multiple sequence alignment. MD simulations evaluated each mutation's effect on insulin stability, using Root Mean Square Deviation (RMSD) for structural shifts and Root Mean Square Fluctuation (RMSF) for flexibility. K-means clustering grouped variants based on these metrics. Among 41 identified INS mutations, several occurred in highly conserved regions, suggesting functional significance. Variants such as T97S, A24V, P52R, L68M and G32S showed increased flexibility, with L68M displaying the highest RMSD, indicating structural destabilization. Based on MD data, we classified mutations as "Unstable," "Flexible," or "Stable." Our findings suggest that structural alterations caused by INS mutations may generate neoantigens, contributing to T1D autoimmunity. This classification provides insight into variant pathogenicity and highlights the importance of conserved regions for insulin function, with potential implications for diagnostics and therapeutics.

Key Words Type 1 Diabetes, INS Gene Mutations, Molecular Dynamics, Insulin Structure, Immunogenicity

INTRODUCTION

Type 1 Diabetes (T1D) is a chronic autoimmune disease characterized by the destruction of insulin-producing β-cells in the pancreas, leading to insulin deficiency and hyperglycemia. While the exact cause remains unclear, both genetic and environmental factors contribute to its development [1].

In 2021, 8.4 million people worldwide had type 1 diabetes, with 18% under 20 years old, 64% between 20-59 and 19% over 60. There were 0.5 million new diagnoses (median age 29) and 35,000 deaths within a year of symptomatic onset. 1.8 million people with type 1 diabetes lived in low-income countries. A 10-year-old diagnosed with type 1 diabetes in 2021 could expect to live 13 years in a low-income country and 65 years in a high-income country. An estimated 3.7 million cases were undiagnosed [2].

Among genetic factors, specific HLA alleles are strongly associated with T1D risk. These genes, involved in immune system regulation, influence T1D susceptibility [1,3]. Beyond HLA genes, mutations in the Insulin Gene (INS) are significant contributors to various diabetes forms, including neonatal diabetes mellitus and Maturity-Onset Diabetes of the Young (MODY) [4]. The INS gene provides instructions for producing insulin, the hormone crucial for regulating blood glucose. Insulin biosynthesis begins with preproinsulin, which is processed into mature insulin [5]. Disruptions to this process, often from INS gene mutations, impair insulin production and secretion, contributing to diabetes [6]. These mutations affect all preproinsulin domains: the signal peptide, B-chain, C-peptide, A-chain and cleavage sites [7].

Untranslated region mutations also contribute to diabetes by disrupting regulatory elements controlling gene expression, impacting insulin mRNA stability or translation and thus

reducing insulin synthesis [8]. These mutations often alter insulin's secondary structure, frequently disrupting disulfide bonds crucial for proper folding. This leads to misfolded proinsulin accumulation in β-cell Endoplasmic Reticulum (ER), triggering inflammation, ER stress and ultimately β-cell death [9].

This has been observed in humans and animal models, like the Akita mouse, whose dominant INS mutation disrupts disulfide bond formation, leading to β-cell death. Studies in NOD mice also show dominant INS mutations can accelerate diabetes development [10].

INS mutations contribute to diabetes through other mechanisms. For example, they can cause hyperproinsulinemia, marked by elevated proinsulin due to impaired proinsulin-to-insulin conversion. Some INS mutations are linked to autoantibody-negative type 1 diabetes, resembling autoimmune T1D but lacking typical autoantibodies, suggesting INS mutations can contribute to diabetes independent of classic autoimmunity [11].

The interplay between INS mutations and the immune system is complex. While their direct immunogenicity in T1D is unclear, they can indirectly influence autoimmunity. Recessive INS mutations, leading to reduced insulin production, may also affect immune cell development and function, potentially increasing autoimmunity risk [12]. INS mutations have diverse structural effects. Signal peptide mutations can disrupt ER targeting and translocation, impairing insulin biosynthesis. Mutations can also affect prohormone convertase cleavage, further impairing insulin processing. Mature insulin sequence mutations can impair Insulin Receptor (IR) binding, leading to insulin resistance and impaired glucose uptake [13].

While research explores therapies for insulin-related disorders, like those targeting β-cell dysfunction and autoimmunity, the unclear pathogenicity of many INS mutations hinders progress [14]. This uncertainty hampers both the understanding of the genetic basis of these disorders and the development of targeted therapies Certain mutations are obviously linked to disease, but others cannot be definitively classified, preventing accurate diagnoses and prognoses. This underscores an important knowledge gap that must be addressed in order to promote personalized medicine [15]. More studies are critically required to systematically define these mutations' functional consequences and clarify their clinical significance, ultimately improving our understanding of insulin biology and the role of insulin in disease [16,17].

However, the structural effects of all known T1D-associated INS mutations are not yet fully understood. Thus, this work seeks to anatomize not only the structural consequences of all T1D-related INS mutations noted in the ClinVar database, but also to do so using canonical molecular modeling and dynamics techniques. By systematically analyzing the structural consequences of these mutations, this study aims to provide a valuable resource for understanding the pathogenesis of T1D and ultimately contribute to the development of personalized therapeutic strategies, where we computationally categorize these mutations which have not been documented previously [18].

METHODS

Identification and Characterization of Missense Variants in the INS Gene

To identify Single Nucleotide Polymorphisms (SNPs), specifically missense variants, within the INS gene, including their associated pathogenicity classifications and corresponding amino acid substitutions, data were retrieved from the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar) [18]. The wild-type amino acid sequence of human insulin (P01308) was obtained from the UniProtKB database (https://www.uniprot.org/uniprotkb). In silico mutagenesis, substituting the wild-type residues with the identified variant amino acids, was performed using Maestro (Schrödinger Release 2023.3). To delineate the domain architecture of preproinsulin and assess the potential impact of the identified mutations on key functional regions, the InterPro database (https://www.ebi.ac.uk/interpro/), a resource encompassing protein families, domains and functional sites, was utilized [19,20].

Conservation Analysis of Insulin Across Multiple Species

To analyze the evolutionary trajectory of insulin across various species, a multiple sequence alignment (MSA) was conducted. Protein sequences for insulin were retrieved from the UniProtKB database for the following species: Homo sapiens (P01308), Sus scrofa (P01315), Bos taurus (P01317), Rattus norvegicus (P01322), Mus musculus (P01325), Cavia porcellus (P01329), Octodon degus (P17715), Gallus gallus (P67970), Danio rerio (O73727), Oryctolagus cuniculus (P01311), Ovis aries (P01318), Canis lupus familiaris (P01321) and Pan troglodytes (P30410).The MSA was performed using Clustal Omega with default parameters, accessed through the European Bioinformatics Institute (EBI) web server (https://www.ebi.ac.uk/Tools/msa/clustalo/) [21]. The resulting alignment was then analyzed for residue conservation using a custom Python script.

Residue conservation was assessed by calculating the conservation score for each position in the alignment. This score represents the frequency of the most common amino acid at that position, expressed as a percentage. Based on these scores, residues were classified into three categories: conserved (scores >80%), semi-conserved (50% ≤ scores ≤ 80%) and non-conserved (scores <50%). To identify potential anomalies or outliers in the conservation pattern, an Isolation Forest model was employed. This unsupervised machine learning algorithm effectively identifies outliers. Trained on conservation scores, the model highlighted anomalous residues. A conservation score of 1 denotes high conservation, reflecting strong amino acid preservation, while 0 indicates no conservation. Protein sequence conservation reflects both ancestral-contemporary amino acid similarity and the probability of evolutionary amino acid changes.

Comparative 3D Structural Analysis of Wild-Type INS and Its Mutants

Following the generation of mutant structures, energy minimization was conducted using the CHARMM forcefield. This minimization process, comprising 100 steps, was executed via OpenMM, a toolkit specifically designed for molecular dynamics simulations and protein model optimization [22]. This step ensured protein structure stability and energetic favorability for subsequent analysis. PyMol was used to visualize mutation-induced structural alterations by mapping protein domains onto their respective insulin regions.

Molecular Dynamics Simulations of Insulin and its Mutants

Molecular Dynamics (MD) simulations were performed using Maestro 12.0 (Schrödinger, LLC) to Investigate Insulin (INS) and mutant stability and dynamics. The protein preparation wizard pre-processed, minimized and dehydrated the protein structures. Salt ions were added and the SPC force field was used as the solvent model. Simulations ran for 50 ns at 300 K. Trajectory analysis, using the interaction diagram module, assessed conformational stability and dynamics, yielding RMSD and RMSF data reflecting structural stability and flexibility. This study analyzed MD simulation data, using INS-Wild as the wild-type reference, to investigate protein structure fluctuations by extracting Alpha-Carbon (CA) atom coordinates from trajectory files. RMSD and RMSF differences were calculated, quantifying each simulation's deviation from the wild-type. K-means clustering (k=3) was performed on standardized RMSD and RMSF values to categorize simulations into "Highest," "Medium," and "Lowest" RMSF groups. Simulations with RMSD and RMSF values exceeding 2 were identified as exhibiting high fluctuations. All analyses were conducted using a custom Python script.

RESULTS

Pathogenicity and Functional Analysis of INS Missense Mutations

Analysis of the ClinVar data revealed a series of missense variants within the INS gene, each with varying implications for pathogenicity. A total of 41 mutations were obtained from ClinVar. Notably, several mutations exhibited a likely pathogenic classification, suggesting a higher probability of contributing to disease development. These mutations include C109F, Y108D, S98I, C95R, L35M, H34P, A24V, P52R and F48C. Further details related to these mutations are given in Table 1.

Another set of mutations was classified as likely risk alleles, indicating a potential association with disease susceptibility. These mutations include Y108C, S101C, C96R, C43G, L35Q, P9R and R6C. The functional consequences of these mutations are not clear, but their identification as potential risk alleles. Several mutations exhibited conflicting classifications of pathogenicity, making it challenging to determine their precise role in disease development. These mutations include Y103C, T97S, C96Y, E93G, S76N, R46Q and A23T. Similarly, a number of mutations also had uncertain significance, meaning their impact on disease development is currently unknown. These mutations include S98C, V92L, G90C, G84R, L68M, G47V, H29D and P9S. Finally, C96S, L35P, G32R and M1I lacked pathogenicity classifications, requiring further research to determine their functional consequences, clarify their disease role and resolve conflicting functional impact classifications. C109F (cysteine to phenylalanine) may disrupt disulfide bonds. Y108D (tyrosine to aspartic acid) alters charge and polarity, potentially affecting protein interactions or stability. S98I (serine to isoleucine) changes polarity, possibly impacting folding. C95R (cysteine to arginine) disrupts potential disulfide bonds with a positive charge. L35M (leucine to methionine) may cause steric clashes. H34P (histidine to proline) may disrupt secondary structure. A24V (alanine to valine) may also cause steric clashes. P52R (proline to arginine) disrupts proline's backbone role. F48C (phenylalanine to cysteine) may lead to aberrant disulfide bond formation. Y108C and S101C (tyrosine/serine to cysteine) introduce sulfhydryl groups, potentially causing inappropriate disulfide bond formation. C96R replaces cysteine with arginine, likely disrupting disulfide bonds. C43G replaces cysteine with glycine, increasing flexibility in the protein backbone and potentially affecting stability. L35Q replaces leucine with glutamine, changing from hydrophobic to polar and possibly affecting protein interactions. P9R replaces proline with arginine, disrupting proline's influence on the protein backbone. Lastly, R6C replaces arginine with cysteine, potentially leading to aberrant disulfide bond formation.

Concentration of Pathogenic Mutations in Conserved Regions of Insulin

Analysis of the MSA of INS with orthologous sequences from other species revealed a high degree of evolutionary conservation, with an overall sequence identity of 77.88%, as shown in Figure 1A-C. This indicates strong selective pressure maintaining the INS protein sequence across species. Despite this overall conservation, a significant proportion of ClinVar-reported mutations with varying functional impacts are located within these highly conserved regions of the INS protein. These mutations include C109F, Y108C, Y108D, S101C, C96S, C96Y, C96R, C95R, V92L, G90C, R89C, L68M, R55C, P52R, F48C, G47V, C43G, L35Q, L35P, L35M, H34P, G32R, G32S, H29D, A24V, A24D, M1I and M1V. In contrast, a smaller subset of mutations is found in non-conserved regions of the INS protein. These include Y103C, S98I, S98C, T97S, E93G, G84R, S76N, R46Q, A23T, P9R, P9S and R6C. The observation of functionally impactful mutations within highly conserved regions suggests that these residues play critical roles in INS function and that even subtle alterations can disrupt protein structure or interactions, leading to altered biological activity.

Table 1: Variants Identified in the INS Gene and Their Predicted Impact

Name	Protein change	Variation ID	Allele ID(s)	Germline classification
c.326G>T (p.Cys109Phe)	C109F	1526011	1517416	Likely pathogenic
c.323A>G (p.Tyr108Cys)	Y108C	21120	33972	Likely risk allele
c.322T>G (p.Tyr108Asp)	Y108D	1526010	1517415	Likely pathogenic
c.308A>G (p.Tyr103Cys)	Y103C	68732	79624	Conflicting classifications of pathogenicity
c.302C>G (p.Ser101Cys)	S101C	68731	79623	Likely risk allele
c.293G>T (p.Ser98Ile)	S98I	1526009	1517414	Likely pathogenic
c.292A>T (p.Ser98Cys)	S98C	435508	429227	Uncertain significance/Uncertain risk allele
c.290C>G (p.Thr97Ser)	T97S	617648	609052	Conflicting classifications of pathogenicity
c.287G>C (p.Cys96Ser)	C96S	68730	79622	not provided
c.287G>A (p.Cys96Tyr)	C96Y	13387	28426	Conflicting classifications of pathogenicity
c.286T>C (p.Cys96Arg)	C96R	918067	906387	Likely risk allele
c.283T>C (p.Cys95Arg)	C95R	3393374	3552450	Likely pathogenic
c.278A>G (p.Glu93Gly)	E93G	393455	380274	Conflicting classifications of pathogenicity
c.274G>T (p.Val92Leu)	V92L	13381	28420	Uncertain significance
c.268G>T (p.Gly90Cys)	G90C	21118	33970	Uncertain significance
c.265C>T (p.Arg89Cys)	R89C	21117	33969	Pathogenic/Likely pathogenic
c.250G>A (p.Gly84Arg)	G84R	68729	79621	Uncertain significance
c.227G>A (p.Ser76Asn)	S76N	1049511	1038051	Conflicting classifications of pathogenicity
c.202C>A (p.Leu68Met)	L68M	68728	79620	Uncertain significance
c.163C>T (p.Arg55Cys)	R55C	13392	28431	Pathogenic/Likely pathogenic/Likely risk allele
c.155C>G (p.Pro52Arg)	P52R	1801850	1859041	Likely pathogenic
c.143T>G (p.Phe48Cys)	F48C	13389	28428	Likely pathogenic/Likely risk allele
c.140G>T (p.Gly47Val)	G47V	21115	33967	Uncertain significance
c.137G>A (p.Arg46Gln)	R46Q	13391	28430	Conflicting classifications of pathogenicity
c.127T>G (p.Cys43Gly)	C43G	21114	33966	Likely risk allele
c.104T>A (p.Leu35Gln)	L35Q	2664354	2831809	Likely risk allele
c.104T>C (p.Leu35Pro)	L35P	68726	79618	not provided
c.103C>A (p.Leu35Met)	L35M	1526013	1517418	Likely pathogenic
c.101A>C (p.His34Pro)	H34P	1526012	1517417	Likely pathogenic
c.94G>C (p.Gly32Arg)	G32R	21123	33975	not provided
c.94G>A (p.Gly32Ser)	G32S	21122	33974	Pathogenic/Likely pathogenic
c.85C>G (p.His29Asp)	H29D	68733	79625	Uncertain significance
c.71C>T (p.Ala24Val)	A24V	36401	45064	Likely pathogenic
c.71C>A (p.Ala24Asp)	A24D	13388	28427	Pathogenic/Likely risk allele
c.67G>A (p.Ala23Thr)	A23T	730224	737916	Conflicting classifications of pathogenicity
c.26C>G (p.Pro9Arg)	P9R	626220	614519	Likely risk allele
c.25C>T (p.Pro9Ser)	P9S	304058	319825	Uncertain significance
c.16C>T (p.Arg6Cys)	R6C	13390	28429	Likely risk allele
c.3G>T (p.Met1Ile)	M1I	65588	76496	not provided
c.3G>A (p.Met1Ile)	M1I	65587	76495	Uncertain significance
c.1A>G (p.Met1Val)	M1V	1455986	1380246	Pathogenic/Likely pathogenic

Numerous mutations in these conserved residues are classified as pathogenic, likely pathogenic, or likely risk alleles, including C109F, Y108C/D, S101C, C95R, P52R, F48C, L35M, H34P, G32S, A24V/D and M1V. While some mutations in conserved regions have uncertain significance (e.g., V92L, G90C, L68M, G47V, H29D and M1I) and others lack definitive classification (C96S, L35P and G32R), the overall trend suggests that alterations within these evolutionarily conserved residues are more likely to disrupt INS function. In contrast, mutations located in non-conserved regions (e.g., Y103C, S98I/C, T97S, E93G, G84R, S76N, R46Q, A23T, P9R/S and R6C) exhibit more varied and often conflicting classifications of pathogenicity, highlighting the challenge of interpreting their functional consequences.

Mutational Analysis of the INS Protein within Functional Domains

Analysis of the INS protein revealed two functional regions: the IlGF_insulin-like domain (cd04367, IPR004825), spanning residues 26-110 and the Insulin conserved site (IPR022353/PS00262), located between residues 95-109. Several mutations fall within these defined regions. Specifically, mutations C109F, Y108C/D, S101C, S98I/C and C95R are located within the conserved insulin site (95-109).

Additionally, mutations T97S, C96S/Y/R, E93G, V92L, G90C, R89C, G84R, S76N, L68M, R55C, P52R, F48C, G47V, R46Q, C43G, L35Q/P/M, H34P, G32R/S, H29D, A24V/D/T and P9R/S are all located within the broader IlGF_insulin_like domain (26-110). The mutations M1I/V and R6C are located outside of both the IlGF_insulin_like domain and the Insulin conserved site, as they are N-terminal to residue 26. The mutations are depicted in Figure 1D whereas the protein domains are depicted in Figure 1E.

Analysis of Insulin Variant Stability and Dynamics

The analysis of the MD simulations revealed several INS protein variants with significantly higher RMSF values

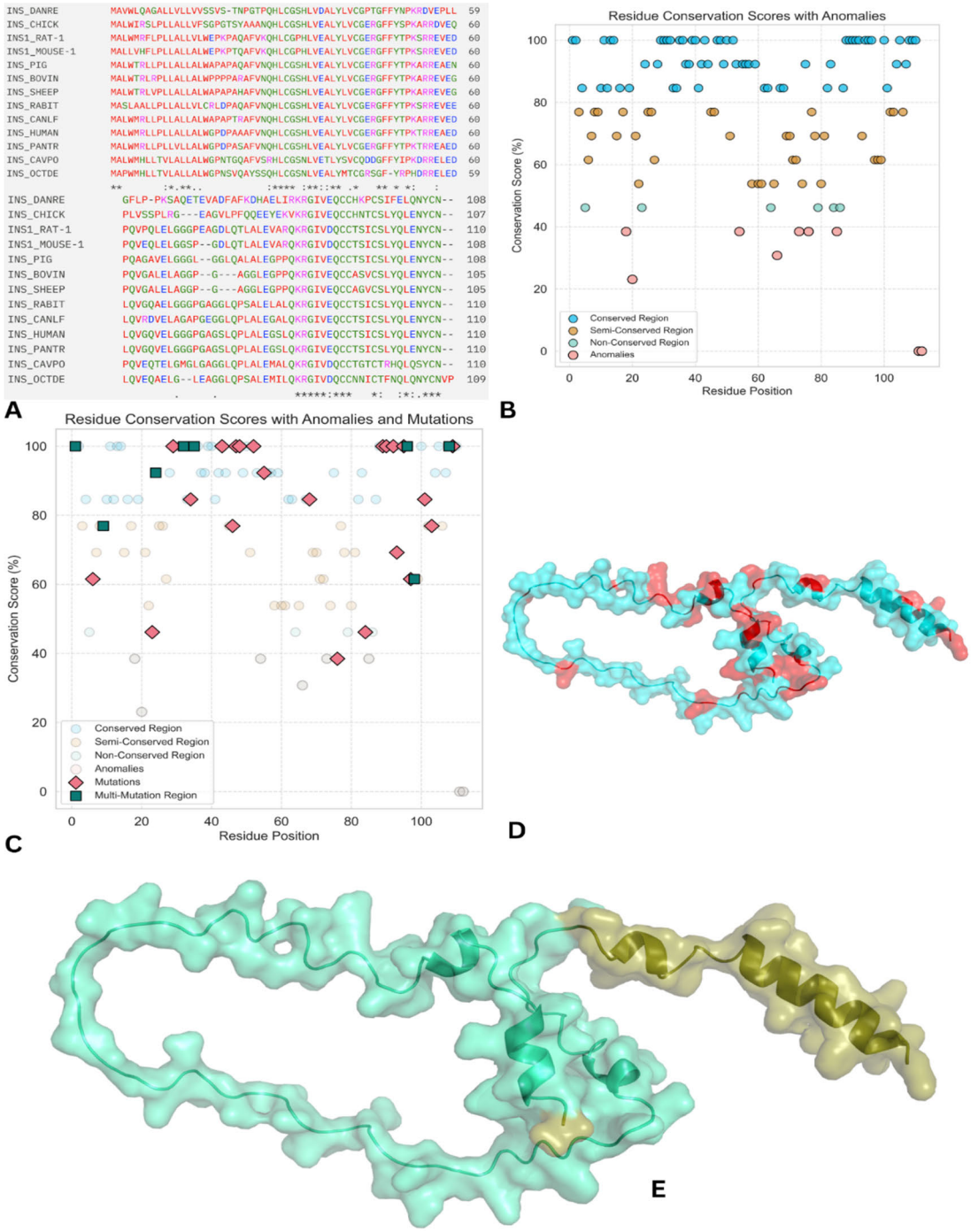

Figure 1: Conservation and location of INS gene mutations. (A-C) MSA and conservation of the INS protein with orthologous sequences from other species, highlighting the high degree of evolutionary conservation. (D) Schematic representation of the INS protein showing the location of ClinVar-reported mutations, colored in red whereas cyan represents the protein. (E) Domain architecture of the INS protein, indicating the IlGF_insulin-like domain and the Insulin conserved site, depicted with light green color

compared to the wild-type INS protein. Analysis of RMSF values relative to the wild revealed significant structural flexibility in several insulin variants. The observation of overall deviation of RMSF from the wild RMSF levels, T97S exhibited the highest RMSF (4.72), indicating dramatic fluctuations potentially due to disrupted hydrogen bonding or altered surface properties from the threonine to serine substitution. A24V also showed substantial fluctuations (RMSF 4.41), likely caused by the alanine to valine substitution disrupting local structure and interactions. Similarly, P52R (RMSF 4.37), with a proline to arginine change, displayed high flexibility, possibly due to disrupted

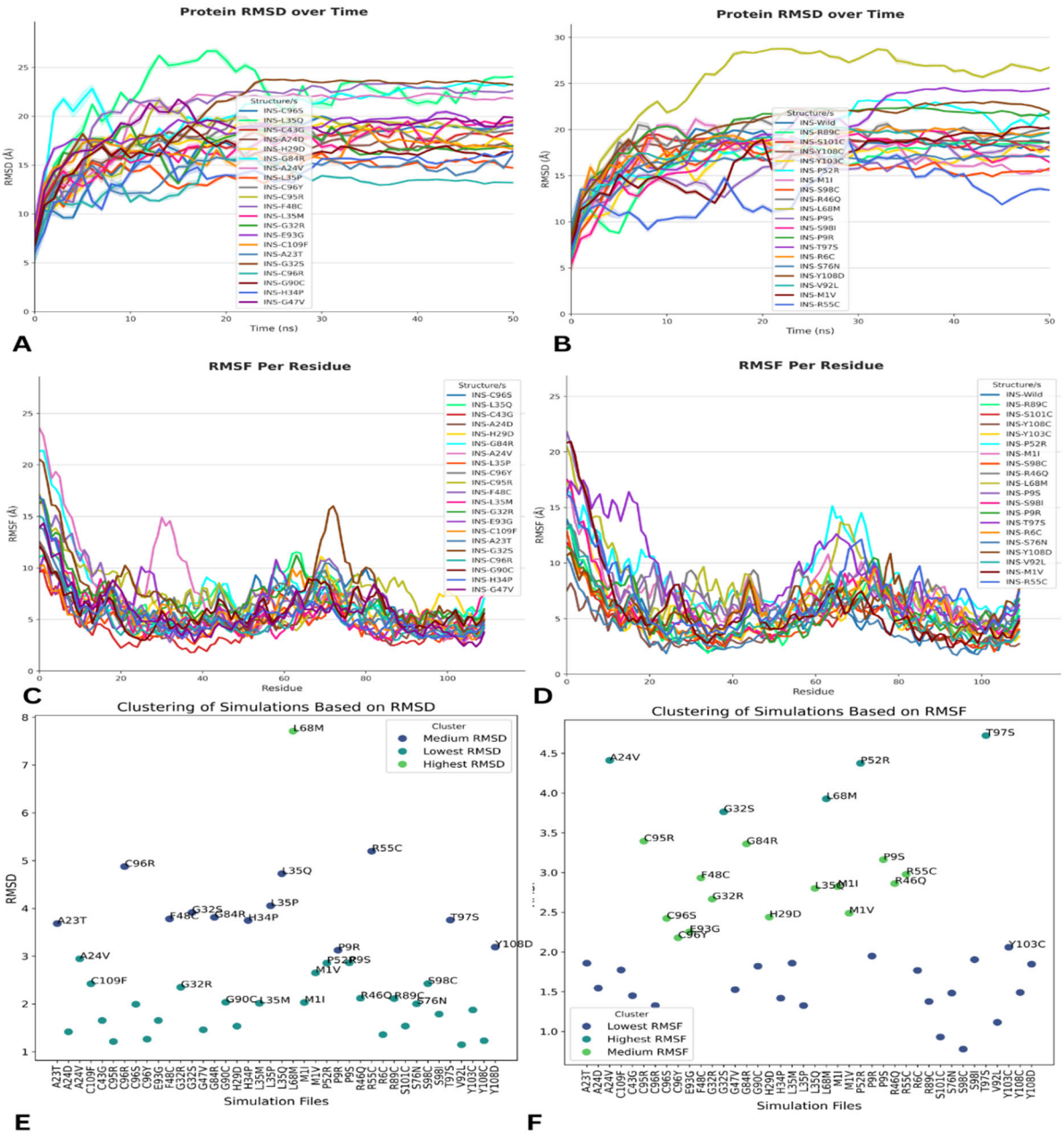

Figure 2: Structural dynamics and stability of INS variants. (A-B) RMSD of the Cα atoms of the INS variants over the course of the 100 ns MD simulations, relative to the starting structure. (C-D) RMSF of the Cα atoms of the INS variants over the course of the 100 ns MD simulations, relative to the wild-type INS protein. (E) K-means clustering of the RMSD values, showing the grouping of variants based on their structural deviation from the wild type. (F) K-means clustering of the RMSF values, showing the grouping of variants based on their flexibility compared to the wild type

secondary structure and new electrostatic interactions. L68M (RMSF 3.93) suggested increased dynamics, likely from altered hydrophobic packing due to the leucine to methionine substitution. Finally, G32S (RMSF 3.77) showed considerable fluctuations, potentially from new hydrogen bonding or steric clashes introduced by the glycine to serine substitution. These elevated RMSF values suggest potential instability or altered functional dynamics in these insulin variants. K-means clustering revealed that these variants deviate significantly from the wild type, exhibiting the "Highest RMSF" deviation, as depicted in Figure 2C-D and 2F.

A cluster of insulin variants exhibited moderately elevated RMSF values, indicating a noticeable increase in structural fluctuations compared to the wild-type protein. Several variants involved cysteine substitutions, potentially disrupting disulfide bonds crucial for stability. C95R (RMSF 3.40), C96S (RMSF 2.42), C96Y (RMSF 2.18) and R55C

Table 2: Classifying Insulin Variants Based on Molecular Dynamics Simulations

Protein	Original Germline Classification	MD Simulation Class
A23T	Conflicting classifications of pathogenicity	Unstable
A24V	Likely pathogenic	Unstable
C109F	Likely pathogenic	Unstable
C96R	Likely risk allele	Unstable
F48C	Likely pathogenic/Likely risk allele	Unstable
G32R	not provided	Unstable
G32S	Pathogenic/Likely pathogenic	Unstable
G84R	Uncertain significance	Unstable
G90C	Uncertain significance	Unstable
H34P	Likely pathogenic	Unstable
L35M	Likely pathogenic	Unstable
L35P	not provided	Unstable
L35Q	Likely risk allele	Unstable
L68M	Uncertain significance	Unstable
M1I	not provided	Unstable
M1V	Pathogenic/Likely pathogenic	Unstable
P52R	Likely pathogenic	Unstable
P9R	Likely risk allele	Unstable
P9S	Uncertain significance	Unstable
R46Q	Conflicting classifications of pathogenicity	Unstable
R55C	Pathogenic/Likely pathogenic/Likely risk allele	Unstable
R89C	Pathogenic/Likely pathogenic	Unstable
S76N	Conflicting classifications of pathogenicity	Unstable
S98C	Uncertain significance/Uncertain risk allele	Unstable
T97S	Conflicting classifications of pathogenicity	Unstable
Y108D	Likely pathogenic	Unstable
A24D	Pathogenic/Likely risk allele	Stable
C43G	Likely risk allele	Stable
C96Y	Conflicting classifications of pathogenicity	Stable
G47V	Uncertain significance	Stable
R6C	Likely risk allele	Stable
V92L	Uncertain significance	Stable
Y108C	Likely risk allele	Stable
C95R	Likely pathogenic	Flexible
C96S	not provided	Flexible
E93G	Conflicting classifications of pathogenicity	Flexible
H29D	Uncertain significance	Flexible
S101C	Likely risk allele	Flexible
S98I	Likely pathogenic	Flexible
Y103C	Conflicting classifications of pathogenicity	Flexible

(RMSF 2.98) all fall into this category, with the arginine substitution in C95R potentially introducing new electrostatic interactions. Glycine to arginine substitutions were also observed, with G32R (RMSF 2.67) and G84R (RMSF 3.36) likely experiencing altered flexibility and new electrostatic interactions. Other variants involved changes in charge or polarity. E93G (RMSF 2.25), H29D (RMSF 2.44), R46Q (RMSF 2.86) and L35Q (RMSF 2.80) all fall into this category, potentially disrupting hydrogen bonding or salt bridges. Changes in hydrophobic character were seen in F48C (RMSF 2.93), M1I (RMSF 2.82) and M1V (RMSF 2.49), possibly affecting protein core packing and N-terminal stability. Finally, P9S (RMSF 3.16) likely experiences altered secondary structure due to the proline to serine change. These moderate increases in RMSF suggest subtle shifts in protein dynamics and potentially altered function.

Lastly, Y103C (RMSF 2.06), located in the Lowest RMSF cluster. A substitution from a tyrosine to a cysteine in protein sequences can disrupt or lead to new aromatic interactions (e.g., from new disulfide bond formation), both of which can affect protein stability and dynamics. The RMSF is a measure of flexibility, so these changes are less pronounced than other variants and it suggests that the overall flexibility of the protein is not giving the same impact.

The comprehensive comparison of individual RMSD detection confirmed few insulin mutants with markedly higher conformational deviation based on overall elevated RMSD relative to wild insulin protein. Among all variants, L68M was the one showing the greatest RMSD (7.71) from the wild-type structure. Conformational changes can be expected due to the substitution of a hydrophobic leucine to a polar methionine, which would result in the mispacking of the core of the protein and changes in the overall structure. Leucine is a branched-chain amino acid, while methionine has a linear side chain with a sulfur atom. This difference in shape and size could disrupt the hydrophobic interactions within the protein core, leading to a significant alteration of the protein's structure. The high RMSD value suggests potential misfolding or a significant alteration of the protein's 3D structure, which could have implications for its function.

A cluster of insulin variants displayed moderately increased RMSD values relative to the wild conformation, indicating noticeable structural deviations compared to the wild type. Several substitutions involving proline, known for its conformational constraints, were observed, including P9R (RMSD 3.13), H34P (RMSD 3.75) and L35P (RMSD 4.06), all potentially disrupting secondary structure. Changes in hydrophobicity and polarity were also common. A23T (RMSD 3.68), F48C (RMSD 3.78), L35Q (RMSD 4.73), T97S (RMSD 3.76) and Y108D (RMSD 3.19) likely experience altered interactions due to these substitutions. Glycine to serine or arginine substitutions, such as G32S (RMSD 3.91) and G84R (RMSD 3.81), could affect backbone flexibility and introduce new electrostatic interactions. Cysteine substitutions, like C96R (RMSD 4.88) and R55C (RMSD 5.20), potentially disrupt disulfide bonds or introduce new ones, significantly impacting conformation. These moderate increases in RMSD suggest noticeable, though not drastic, structural deviations from the wild-type insulin structure.

This cluster encompasses variants exhibiting the lowest RMSD values, all below 3 Å, among those exceeding the 2 Å threshold. While these variants do show some structural deviations from the wild type, their RMSD values suggest relatively minor changes in overall conformation.

Specifically, the A24V (2.94 Å), C109F (2.42 Å), G32R (2.35 Å), G90C (2.04 Å), L35M (2.02 Å), M1I (2.03 Å), M1V (2.65 Å), P52R (2.86 Å), P9S (2.86 Å), R46Q (2.12 Å), R89C (2.11 Å), S76N (2.00 Å) and S98C (2.43 Å) substitutions, while potentially affecting local structure, packing, hydrophobic interactions, electrostatic interactions, disulfide bond formation, or secondary structure, appear to induce only limited overall structural changes, as depicted in Figure 2A-B and 2E.

Finally when compared to the original germline classification, we propose a new MD-simulations based classification of the INS variants, as given in Table 2.

DISCUSSION

This study investigated the pathogenicity and functional consequences of missense mutations within the INS gene, focusing on their impact on insulin protein stability and dynamics. Our analysis combined data from ClinVar,

sequence conservation analysis, functional domain mapping and MD simulations to provide a comprehensive assessment of these variants.

ClinVar data revealed a spectrum of classifications for the identified INS mutations, ranging from clearly pathogenic to uncertain significance, with many falling into ambiguous or conflicting categories. This highlights the ongoing challenge of accurately predicting the clinical impact of missense variants, particularly in genes like INS where subtle alterations can have profound physiological consequences. Our study aimed to address this challenge by integrating computational approaches to gain deeper insights into the functional effects of these mutations.

Two recent studies have revealed significant challenges in the standardized classification of human genetic variants. One large-scale genomic analysis, comparing variant frequencies with disease prevalence across numerous genes and conditions, found substantial inflation of pathogenic variants, particularly among those with weaker supporting evidence and rare variants, suggesting widespread misclassification. This inflation was replicated in a separate analysis of endocrine tumor syndromes using ClinVar data, which also showed inflated genetic risk. While these findings highlight the problem of overestimating pathogenicity in genetic databases, the larger genomic study also underscores the crucial role of resources like ClinVar in facilitating comparison and validation of variant classifications, ultimately leading to progressive improvements in accuracy over time [16,17,22].

The analysis of evolutionary conservation indicated the critical importance of specific residues within the insulin protein. The high degree of sequence identity observed across species emphasizes the strong selective pressure maintaining INS function. Intriguingly, a substantial proportion of the ClinVar-reported mutations, including many classified as pathogenic or likely pathogenic, are located within these highly conserved regions. This observation strongly suggests that these conserved residues play crucial roles in insulin's structure, stability and interactions and that even seemingly minor amino acid substitutions can disrupt these critical functions. Conversely, mutations in non-conserved regions are often presented with conflicting or uncertain pathogenicity classifications, suggesting that while these changes might have some impact, their effects are likely more subtle and difficult to predict based on sequence analysis alone.

Identifying the mutations in known functional domains of insulin allowed us to strengthen the role of specific areas. These findings underscore the functional importance of both the Insulin conserved site and the wider IlGF_insulin-like domain as they contain a higher concentration of pathogenic and likely pathogenic mutations. Mutations in these domains are more likely to directly impact insulin binding to its receptor, folding, or stability. The prevalence of pathogenic mutations across conserved regions of insulin suggests that these residues serve an essential role in its proper folding, stability and activity [23].

These regions often include amino acids involved in disulfide bond formation, receptor binding and maintaining the protein's overall three-dimensional structure. Mutations in these areas are more likely to be disruptive and result in disease. Mutations in non-conserved regions of insulin are often classified with conflicting or uncertain pathogenicity because these regions are more tolerant of sequence variation without disrupting the protein's core function. These regions may not directly participate in key structural or functional interactions, so changes there may have subtle or no effects, making it difficult to predict their clinical significance [24].

Key functional domains of insulin include the signal peptide, essential for directing insulin to the endoplasmic reticulum for processing and secretion; the A-chain and B-chain, which form the core of mature insulin and are crucial for receptor binding and biological activity; the C-peptide, which plays a role in proinsulin folding and is used as a proxy to measure insulin production; disulfide bonds, critical for stabilizing the structure of the insulin protein; and receptor binding sites, specific regions within the A- and B-chains that interact with the insulin receptor [25]

Mutations in the conserved sites of insulin are likely to disrupt the function of insulin, with the most frequent mutations associated with neonatal diabetes and other forms of diabetes often found in these conserved regions of the A and B chains, including the signal peptide. Mutations in the IGF_insulin-like domain may have effects on function, but these may be subtle since there is more divergence in the IGF_insulin-like domains between species [26].

Even minor amino acid substitutions in conserved regions of insulin can have significant consequences, disrupting protein folding, receptor binding and protein stability. The MD simulations provided valuable insights into the dynamic behavior of the mutant insulin proteins. RMSF analysis revealed significant increases in flexibility for several variants, suggesting potential instability or altered functional dynamics. The variants T97S, A24V, P52R, L68M and G32S exhibited the highest RMSF values, indicating substantial fluctuations that could disrupt crucial interactions or lead to misfolding. These findings align with the observed location of these mutations within conserved regions and functional domains, reinforcing the idea that these changes disrupt critical structural elements. Similarly, variants like C95R, C96S/Y, R55C, G32R, G84R and others showed moderately elevated RMSF, suggesting more subtle but still potentially impactful changes in protein dynamics. Interestingly, Y103C, despite its location in a conserved region, showed relatively low RMSF, highlighting the complexity of predicting functional impact based solely on conservation.

RMSD analysis further corroborated the instability observed in several variants. L68M, with the highest RMSD, displayed a substantial deviation from the wild-type structure, indicating a significant alteration in overall conformation. Other variants, including P9R, H34P, L35P, A23T, F48C, L35Q, T97S, Y108D, G32S, G84R, C96R and R55C, also exhibited moderately increased RMSD, suggesting noticeable

structural deviations. These findings, combined with the RMSF data, provide strong evidence that these mutations can destabilize the insulin protein and potentially impair its function.

Therefore, here we provide a revised categorization of the INS variants by linking the data from ClinVar, conservation analysis, functional domain mapping and MD simulations. Based on our MD-based classification schemes classifies variants as either "Unstable," "Flexible," or "Stable" according to their simulated dynamics. This classification provides a functional context for interpreting the often ambiguous or conflicting germline classifications for this variant present in ClinVar. Several previously classified variants of uncertain significance or given conflicting pathogenicity assignments were marked as "Unstable" via our MD simulations and were determined to potentially impact function in previously un-delineated ways deserving further exploration. [27] Likewise, differences that fall under the category of “Flexible” might have minor instabilities in their dynamics, that could change the activity of insulin or some interactions related to it, even if they do not lead to gross structural instability.

We show that our findings suggest that INS gene mutations may lead to significant structural changes in the insulin protein. These changes, as demonstrated by the RMSF and RMSD analyses, can create "neoantigens" - new protein forms that the immune system recognizes as foreign. The unstable and flexible variants identified in the MD simulations are more likely to be processed and presented by antigen-presenting cells, potentially initiating or exacerbating the autoimmune response. Essentially, the structural deviations caused by the mutations can break immune tolerance to insulin [28,29].

Findings emphasize the concentration of pathogenic mutations within the Insulin conserved site and the broader IGF_insulin-like domain. These regions are crucial for insulin function, but they are also likely targets for immune recognition. Mutations in these conserved areas not only affect insulin's biological activity but also increase the likelihood of generating T cell epitopes (regions recognized by T cells) that drive the autoimmune attack. The fact that these regions are highly conserved suggests that even subtle alterations can be perceived as "non-self" by the immune system. [12,15,30,31,32].

This study demonstrates the by combining computational and bioinformatic approaches to investigate the functional consequences of missense mutations. Our findings provide valuable insights into the mechanisms by which INS mutations can affect insulin stability and dynamics, ultimately contributing to altered biological activity. While our study focuses on the biophysical effects of these mutations, future work should investigate their impact on insulin signaling, receptor binding and downstream physiological processes. Furthermore, functional studies, such as in vitro assays of insulin activity and in vivo models, are necessary to validate our computational predictions and fully elucidate the clinical relevance of these INS variants. This integrated approach will be crucial for improving the diagnosis and management of diabetes-related disorders associated with INS gene mutations.

CONCLUSION

This study demonstrates that missense mutations in the INS gene can significantly alter insulin protein structure and dynamics, potentially creating neoantigens and contributing to the immunogenicity of T1D. MD simulations revealed that certain variants exhibit increased flexibility and instability compared to wild-type insulin, particularly those located within conserved regions like the Insulin conserved site and the IGF insulin-like domain. These structural changes can affect insulin function and increase the likelihood of immune recognition, potentially breaking immune tolerance and driving the autoimmune attack in T1D. The MD-based classification of variants as "Unstable," "Flexible," or "Stable" provides functional context for ClinVar classifications and highlights the importance of further research, including in vitro and in vivo studies, to validate these findings and translate them into improved diagnostics and personalized therapies for T1D.

Acknowledgement

The authors would like to express their sincere thanks and appreciation to the Department of Biology, College of Education for Pure Science, University of Diyala, Iraq, for their valuable support and assistance throughout the course of this study.

REFERENCES

[1] Redondo, Maria J. and Noel G. Morgan. “Heterogeneity and endotypes in type 1 diabetes mellitus.” *Nature Reviews Endocrinology,* vol. 19, no. 9, June 2023, pp. 542-554. http://dx.doi.org/10.1038/s41574-023-00853-0.

[2] Gregory, Gabriel A. *et al.* “Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: A modelling study.” *The Lancet Diabetes & Endocrinology,* vol. 10, no. 10, October 2022, pp. 741-760. http://dx.doi.org/10.1016/s2213-8587(22)00218-2.

[3] Scholten, Bernt Johan von *et al.* “Current and future therapies for type 1 diabetes.” *Diabetologia,* vol. 64, no. 5, February 2021, pp. 1037-1048. http://dx.doi.org/10.1007/s00125-021-05398-3.

[4] Roep, Bart O. *et al.* “Type 1 diabetes mellitus as a disease of the β-cell (do not blame the immune system?).” *Nature Reviews Endocrinology,* vol. 17, no. 3, December 2020, pp. 150-161. http://dx.doi.org/10.1038/s41574-020-00443-4.

[5] R. David Leslie *et al.* “Adult-Onset Type 1 Diabetes: Current Understanding and Challenges.” *Diabetes Care* vol. 44, no. 11, December 2020, pp. 2449-2456. https://pubmed.ncbi.nlm.nih.gov/35831242/.

[6] Rahmati, Masoud *et al.* “The global impact of COVID-19 pandemic on the incidence of pediatric new-onset type 1 diabetes and ketoacidosis: A systematic review and meta-analysis.” *Journal of Medical Virology,* vol. 94, no. 11, July 2022, pp. 5112-5127. http://dx.doi.org/10.1002/jmv.27996.

[7] ahaya, Tajudeen and Titilola Salisu. “Genes predisposing to type 1 diabetes mellitus and pathophysiology: A narrative review.” *Medical Journal of Indonesia,* vol. 29, no. 1, March 2020, pp. 100-9. http://dx.doi.org/10.13181/mji.rev.203732.

[8] Kushi, Ryo, *et al.* "Insulin resistance and exaggerated insulin sensitivity triggered by single-gene mutations in the insulin signaling pathway." *Diabetology International,* vol. 12, no. 1, July 2020. http://dx.doi.org/10.1007/s13340-020-00455-5.

[9] Margaritis, Kosmas *et al.* "Micro-rna implications in type-1 diabetes mellitus: A review of literature." *International Journal of Molecular Sciences,* vol. 22, no. 22, November 2021, pp. 12165-0. http://dx.doi.org/10.3390/ijms222212165.

[10] Akil, Ammira Al Shabeeb *et al.* "Diagnosis and treatment of type 1 diabetes at the dawn of the personalized medicine era." *Journal of Translational Medicine,* vol. 19, no. 1, April 2021. http://dx.doi.org/10.1186/s12967-021-02778-6.

[11] Haichen Zhang *et al.* "Monogenic Diabetes: A Gateway to Precision Medicine in Diabetes." *J. Clin. Invest.* vol. 131, no. 3, February 2021. https://pubmed.ncbi.nlm.nih.gov/33529164/

[12] Al-Beltagi, Mohammed *et al.* "Insulin-resistance in paediatric age: Its magnitude and implications." *World Journal of Diabetes,* vol. 13, no. 4, April 2022, pp. 282-307. http://dx.doi.org/10.4239/wjd.v13.i4.282.

[13] Ramasubbu, Kanagavalli and V. Devi Rajeswari. "Impairment of insulin signaling pathway pi3k/akt/mtor and insulin resistance induced ages on diabetes mellitus and neurodegenerative diseases: A perspective review." *Molecular and Cellular Biochemistry,* vol. 478, no. 6, October 2022, pp. 1307-1324. http://dx.doi.org/10.1007/s11010-022-04587-x.

[14] Støy, Julie *et al.* "In celebration of a century with insulin – update of insulin gene mutations in diabetes." *Molecular Metabolism,* vol. 52, October 2021, pp. 101280-0. http://dx.doi.org/10.1016/j.molmet.2021.101280.

[15] Low, Blaise Su Jun *et al.* "Decreased glut2 and glucose uptake contribute to insulin secretion defects in mody3/hnf1a hipsc-derived mutant β cells." *Nature Communications,* vol. 12, no. 1, May 2021, pp. 3133. http://dx.doi.org/10.1038/s41467-021-22843-4.

[16] Petkovic, Sonja and Katja Lohmann. "Disease-causing or benign? challenges in genetic variant interpretation and limitations of clinvar." *Movement Disorders,* vol. 33, no. 8, August 2018, pp. 1271. http://dx.doi.org/10.1002/mds.94.

[17] Toledo, Rodrigo, A. "Inflated pathogenic variant profiles in the clinvar database." *Nature Reviews Endocrinology,* vol. 14, no. 7, July 2018, pp. 387-389. http://dx.doi.org/10.1038/s41574-018-0034-0.

[18] Landrum, Melissa J. *et al.* "Clinvar: Public archive of interpretations of clinically relevant variants." *Nucleic Acids Research,* vol. 44, no. D1, November 2015, pp. D862-D868. http://dx.doi.org/10.1093/nar/gkv1222.

[19] Paysan-Lafosse, Typhaine, *et al.* "Interpro in 2022." *Nucleic Acids Research,* vol. 51, no. D1, November 2022, pp. D418-D427. http://dx.doi.org/10.1093/nar/gkac993.

[20] UniProt Consortium, "Uniprot: A hub for protein information." *Nucleic Acids Research,* vol. 43, no. D1, January 2015, pp. D204-D212. http://dx.doi.org/10.1093/nar/gku989.

[21] Sievers, Fabian, and Desmond G. Higgins. "Clustal omega, accurate alignment of very large numbers of sequences." Methods in Molecular Biology, vol. 1079, August 2014, pp. 105-116. http://dx.doi.org/10.1007/978-1-62703-646-7_6.

[22] Peter Eastman *et al.* "OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials." *The Journal of Physical Chemistry B* vol. 128, no. 1, December 2023, pp. 109-116. https://doi.org/10.1021/acs.jpcb.3c06662.

[23] Shah, Naisha, *et al.* "Identification of misclassified clinvar variants via disease population prevalence." *The American Journal of Human Genetics,* vol. 102, no. 4, April 2018, pp. 609-619. http://dx.doi.org/10.1016/j.ajhg.2018.02.019.

[24] Cook, Taylor W., *et al.* "Understanding insulin in the age of precision medicine and big data: Under-explored nature of genomics." *Biomolecules,* vol. 13, no. 2, January 2023, pp. 257-0. http://dx.doi.org/10.3390/biom13020257.

[25] Dhayalan, Balamurugan, *et al.* "Structural lessons from the mutant proinsulin syndrome." *Frontiers in Endocrinology,* vol. 12, September 2021, pp. 754693-754693. http://dx.doi.org/10.3389/fendo.2021.754693.

[26] White, Morris F., and C. Ronald Kahn. "Insulin action at a molecular level–100 years of progress." *Molecular Metabolism,* vol. 52, October 2021, pp. 101304-0. http://dx.doi.org/10.1016/j.molmet.2021.101304.

[27] Geng, Hao *et al.* "Applications of molecular dynamics simulation in structure prediction of peptides and proteins." *Computational and Structural Biotechnology Journal,* vol. 17, July 2019, pp. 1162-1170. http://dx.doi.org/10.1016/j.csbj.2019.07.010.

[28] Childers, Matthew Carter and Valerie Daggett. "Insights from molecular dynamics simulations for computational protein design." *Molecular Systems Design & Engineering,* vol. 2, no. 1, February 2017, pp. 9-33. http://dx.doi.org/10.1039/c6me00083e

[29] Rognan, Didier. "Molecular dynamics simulations: A tool for drug design." In: 3D QSAR in Drug Design. Three-Dimensional Quantitative Structure Activity Relationships, Kubinyi, H., Folkers, G. and Y.C. Martin (Eds.)., Dordrecht, Springer Netherlands, 2002, pp. 181-209. http://dx.doi.org/10.1007/0-306-46857-3_11.

[30] Rowe, Glenn C., *et al.* "Increased energy expenditure and insulin sensitivity in the high bone mass δfosb transgenic mice." *Endocrinology,* vol. 150, no. 1, September 2008, pp. 135-143. http://dx.doi.org/10.1210/en.2008-0678.

[31] Zhang, Xing, *et al.* "Depletion of junb increases adipocyte thermogenic capacity and ameliorates diet-induced insulin resistance." *Nature Metabolism,* vol. 6, no. 1, January 2024, pp. 78-93. http://dx.doi.org/10.1038/s42255-023-00945-1.

[32] Lee, Hanseul and Seung Jae V. Lee. "Recent progress in regulation of aging by insulin/igf-1 signaling in caenorhabditis elegans." *Molecules and Cells,* vol. 45, no. 11, November 2022, pp. 763-770. http://dx.doi.org/10.14348/molcells.2022.0097.